FORECAST ERROR REDUCTION BY PREPROCESSED HIGH-PERFORMANCE STRUCTURAL BREAK DETECTION

Dirk Pauli, Jens Feller, Bernhard Mauersberg
FCE Frankfurt Consulting Engineers GmbH, Frankfurter Str.5, 65239 Hochheim/Main, Germany

Ingo J. Timm
Department IV, Institute of Business Information Systems, Business Informatics I, University of Trier, 54286 Trier, Germany

Keywords: Change-point detection, Hypothesis testing, Chernoff bounds, Binomial distribution, Additive changes, Non-additive changes, Multiple structural break detection.

Abstract: In this paper a new method for detecting multiple structural breaks, i.e. undesired changes of signal behavior, is presented and applied to real-world data. It will be shown how Chernoff Bounds can be used for high-performance hypothesis testing after preprocessing arbitrary time series to binary random variables using k-means-clustering. Theoretical results from part one of this paper have been applied to real-world time series from a pharmaceutical wholesaler and show striking improvement in terms of forecast error reduction, thereby greatly improving forecast quality. In order to test the effect of structural break detection on forecast quality, state of the art forecast algorithms have been applied to time series with and without previous application of structural break detection methods.

1 INTRODUCTION

Structural break detection concentrates on discovering time points at which properties of time series change significantly. This term is used e.g. in (Perron, 2006) but other terms like change-point, event, novelty, anomaly or abnormality detection, e.g. in (Kawahara and Sugiyama, 2009), (Markou and Singh, 2003), (Guralnik and Srivastava, 1999), (Ma and Perkins, 2003), and (Ibaida et al., 2010) refer to this problem in a similar manner. The problem itself varies with its application. For example, consider a time series with a stable trend. Despite changes in statistical moments of the distribution, this results in no loss of forecast quality if the data represents the historical demand of an article and the task is to forecast future demands. In contrast, if the same time series is a vibration signal of a gas turbine, a stable trend leads to an undesired state of the machine and must be detected as soon as possible, compare (Feller et al., 2010). Further real-world applications include e.g. fraud detection in (Murad and Pinkas, 1999), anomaly detection for spacecraft in (Fujimaki et al., 2005) or (Schwabacher et al., 2007), detecting abnormal driving conditions in (Gustafsson, 1998), and anomaly detection in multi-node computer systems in (Ide and Kashima, 2004) to name but a few. More application domains and examples are provided in (Chandola et al., 2009). All these applications emphasize the importance and need of algorithms for change-point detection for a broad community.

Another real-world application is the forecast of future demands, which is a crucial element of calculating an optimal stock policy. In many cases, large amounts of data are available, but information cannot be retrieved completely, due to limited resources in terms of e.g. computing time or inefficient algorithms. In order to gain the full information available an automated, reliable, and efficient work flow has to be established.

In this paper, a novel approach to structural break detection is introduced in order to reduce forecast errors and thereby increase accuracy and reliability of forecasts. The algorithm is validated on a real-world data set consisting of 8002 independent time series of historical demands of articles of a pharmaceutical wholesaler. The performance of the new algorithm is measured in terms of forecast error reduction, statistical power, significance and runtime on this particular data set.

Pauli D., Feller J., Mauersberg B. and J. Timm I..
FORECAST ERROR REDUCTION BY PREPROCESSED HIGH-PERFORMANCE STRUCTURAL BREAK DETECTION.
DOI: 10.5220/0003457202620271
In *Proceedings of the 8th International Conference on Informatics in Control, Automation and Robotics* (ICINCO-2011), pages 262-271
ISBN: 978-989-8425-74-4

Related work, compare (Basseville and Nikiforov, 1993), shows that a common approach in the area of change-point detection is to divide the task at least into two parts: the first step generates residuals of the original measurements that reflect the changes of interest, e.g. the residuals are close to zero before and nonzero after the change. The second step contains the design of a decision rule based upon these residuals. The algorithm presented in this paper proceeds in a similar way. The first task is to transform an arbitrary time series $x_1, \ldots, x_s \in \mathbb{R}$ to a sequence of binary numbers, which is interpreted as the outcome of a binary stochastic process $\{Y_i\}_{i \in \mathbb{N}}$ with $\Omega := \{0,1\}$. Afterwards Chernoff Inequalities are used for hypothesis testing, i.e. to estimate the probability of subsequences and detect structural breaks.

In the context of this paper, the new algorithm is adjusted to detect additive changes. However, the novel approach can be adapted to detect nonadditive changes as well, which is discussed in section 2.4. In e.g. (Basseville and Nikiforov, 1993) additive changes are defined as shifts in the mean value of a signal, while nonadditive changes are defined as changes in variance, correlations, spectral characteristics, or dynamics of the signal or system. Both definitions will be used throughout this paper. Furthermore, this paper concentrates on offline detection, since it is sufficient for the current application. An online variant of this algorithm will be discussed in section 4.

The paper is structured as follows: section 2.1 presents how Chernoff's Inequalities can be used for high-performance hypothesis testing to detect structural breaks. Section 2.2 provides the design of a transformation routine that fulfils the goal of the case study and reflects additive changes. In section 2.3 the basic algorithm is extended to multiple structural break detection. In order to show the flexibility of the novel approach, the detection of nonadditive changes is discussed in section 2.4. Section 3 contains the application of the new algorithm to a real-world problem. Since forecast error reduction will be used as a key performance indicator of the new algorithm, a set of forecast methods is shortly introduced in section 3.1. Test scenarios and error estimates are defined in section 3.2. The results of the case study and performance indicators of the algorithm are presented in section 3.3. In section 4 results of this paper are discussed and potential future enhancements are suggested.

2 A NOVEL APPROACH TO HIGH-PERFORMANCE STRUCTURAL BREAK DETECTION

The algorithm used in this paper can be separated into two parts. The first step is to generate random variables $y_i \in \{0,1\}$ for all $i \in [1,s]$ from the corresponding x_i in order to satisfy the requirements of the variant of Chernoff's bounding method used here. The second step is to prepare and to perform a hypothesis test. The authors of this paper decided to start with step two for reasons of clarity, therefore it will be assumed until section 2.2 that a routine $P : \mathbb{R} \to \{0,1\}$ does exist to transform x_i adequately.

2.1 Chernoff's Bounding Method for Hypothesis Testing

In this section the application of Chernoff's bounding method to detect structural breaks in time series y_i is presented. First Chernoff's Inequality is described.

Theorem (Chernoff's Inequality). Given s independent Bernoulli-experiments $y_1, \ldots, y_s$ with probability $Pr[y_i = 1] = p$ and $Pr[y_i = 0] = 1 - p$, then for each $\alpha > 0$

$$Pr\left[\sum_{i=1}^{s} y_i \geq (1+\alpha) \cdot p \cdot s\right] \leq e^{-\frac{\alpha^2 \cdot p \cdot s}{3}} \tag{1}$$

and for each $\alpha \in [0,1]$

$$Pr\left[\sum_{i=1}^{s} y_i \leq (1-\alpha) \cdot p \cdot s\right] \leq e^{-\frac{\alpha^2 \cdot p \cdot s}{2}} \tag{2}$$

holds, compare (Chernoff, 1952).

In other words, large linear deviations from the expectation are highly improbable. Starting at this, point a hypothesis test can be defined as follows: it is assumed that all y_i are independent and identically distributed, therefore the sum of events y_i will only exceed each bound with probability less than γ, where

$$\gamma = e^{-\frac{\alpha^2 \cdot p \cdot s}{c}} \tag{3}$$

and $c \in \{2,3\}$. If bounds are exceeded, the assumption is considered to be wrong and the hypothesis is rejected. The probability γ is antiproportional to the risk of making a wrong decision.

The central idea of this paper is to perform hypothesis tests for each continuous subsequence of length τ and verify whether the occurrence of events $y_i = 1$ notably differ from their expectation. If they

do, the distribution of y_i has changed or differs between certain subsequences and a structural break is considered. As the distribution of 1's and 0's is assumed to be binomial, p can be estimated as follows:

$$A^0 = \{j | j \in \{1,\ldots,s\}, y_j = 0\} \quad (4)$$

$$A^1 = \{j | j \in \{1,\ldots,s\}, y_j = 1\} \quad (5)$$

Then $r^0 := \frac{|A^0|}{s}$ and $r^1 := \frac{|A^1|}{s}$ lead to $p := r^1$.

The upper and lower bounds are dependent on α_u and α_l, which can be estimated for a given $\gamma \in (0,1)$ as follows:

$$\gamma = e^{-\frac{\alpha^2 \cdot p \cdot \tau}{3}} \Leftrightarrow \alpha_u = \sqrt{-\frac{3 \cdot ln(\gamma)}{\tau \cdot p}} \quad (6)$$

$$\gamma = e^{-\frac{\alpha^2 \cdot p \cdot \tau}{2}} \Leftrightarrow \alpha_l = \sqrt{-\frac{2 \cdot ln(\gamma)}{\tau \cdot p}} \quad (7)$$

The next step is to test $\forall i \in [1,\ldots,s-\tau+1]$, whether the distribution of $y_i,\ldots,y_{i+\tau-1}$ is likely using Chernoff bounds for an estimated p. In other words, it is checked if the sum over $y_i,\ldots,y_{i+\tau-1}$ deviates from its expectation by more than a factor of $1+\alpha$ or $1-\alpha$, respectively. Such a deviation of the sum from its expectation can only happen with a probability less than or equal to γ. As gamma is small, deviation leads to the hypothesis being rejected, and a structural break is assumed. If the 1's are uniformly distributed according to p, the hypothesis will hold with probability $1-\gamma$.

Hypothesis H_i $\forall i \in [1,\ldots,s-\tau+1]$ is tested and set as follows:

$$H_i = \begin{cases} \text{reject} & \text{if } (S \geq (1+\alpha_u) \cdot p \cdot \tau) \\ & \vee (S \leq (1-\alpha_l) \cdot p \cdot \tau) \\ \text{accept} & \text{else} \end{cases} \quad (8)$$

with $S := \sum_{j=i}^{i+\tau-1} y_j$.

If at least for one sequence $y_i,\ldots,y_{i+\tau-1}$ the hypothesis H_i is rejected, then a clustering of 1's or 0's can be assumed and a structural break is likely. If a structural break occurs, it is valuable to know the exact time index of the break, e.g. to cut off the time series to improve forecasting methods.

Case I: Actual Samples belong to the Group of 0's. Select rejected hypothesis k with smallest index, which means that sequence $y_k,\ldots,y_{k+\tau-1}$ is assumed to be unlikely. Returning $b = k$ as the result of the analysis might cause a loss of reliable samples in the time series. Therefore return

$$b = min\{j | j \in \{k,\ldots,k+\tau-1\}, y_j = 1\} \quad (9)$$

as the first index of the invalid subsequence.

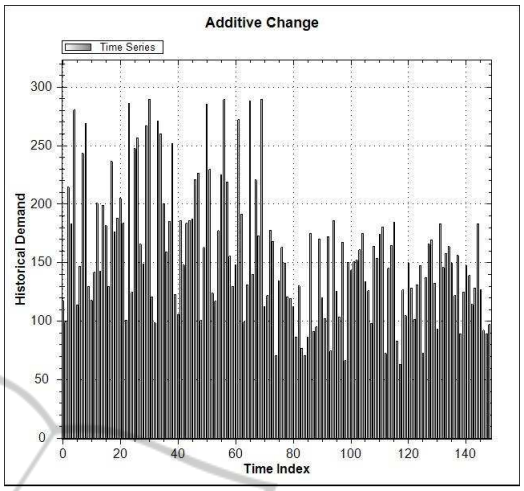

Figure 1: Example of an additive change. At break time index 69 the arithmetic mean shifts from 181 items to 119 items.

Case II: Actual Samples belong to the Group of 1's. Select holding hypothesis k with smallest index. In contrast to case I, the sequence $y_k,\ldots,y_{k+\tau-1}$ is assumed to be likely and to keep as many samples as possible, return

$$b = min\{j | j \in \{k,\ldots,k+\tau-1\}, y_j = 0\} \quad (10)$$

as the first index of the invalid subsequence.

2.2 Transformation Routine

Finding an adequate transformation routine of course requires a clear definition of what shall be detected as a discontinuous behavior, and consequently its design is absolutely dependent on this definition. Therefore this paper cannot provide a general answer to this problem. Instead, a strategy for the practical problem considered in the context of this paper is discussed in this section.

Consider a company with a large amount of products, whose demand needs to be forecasted day by day, e.g. a supermarket or any wholesaler. Obviously, forecasting cannot be done manually in such cases, and reliable strategies have to be chosen to solve the problem. The success of forecasting strategies depends on the quality of considered time series and on the robustness of applied methods. Having a large amount of articles and the necessity of daily forecasts multiplies to a number of events, which is likely to bring up even rare cases. However, even the most robust strategies cannot cover each and every situation. Hence a different approach to improving the forecast quality is to improve the quality of the input data using preprocessing methods, and especially in this con-

text a method to detect and remove structural breaks. Following a structural break means a rapid and strong shift of the mean demand of a certain article. In other words, one can find two different distributions which can be separated at a certain point in time. Figure 1 shows an example for such a strong and rapid shift of the mean. One can see that in week 69 the behavior of the time series changes dramatically. The mean demand changes from 181 items based on the deliveries until week 69 to 119 items based on the deliveries starting from week 70. Since safety stock levels are often affected by variance or standard deviation, an estimation of stock level based on the complete time series can lead to overstocking in cases as described above.

Taking the previous considerations into account, the task can be summarized as follows: If a set of data is likely to correspond to two different distributions, is there a point in time which can be used to differentiate between both distributions, or do the random numbers come alternating from both distributions? Having obtained the results from section 2.1 it is necessary to find an adequate transformation routine.

A well-known clustering algorithm is the *k*-means-clustering as described e.g. in (Press et al., 2007). Clustering is known to be NP-hard in standard scenarios, hence polynomial clustering heuristics like *k*-means-clustering do not guarantee optimal solutions. Since in this case clustering is performed for only one dimension the algorithm converges to the optimum as described in (Hartigan and Wong, 1979). The goal of the algorithm is to find k clusters in n-dimensional space, where a cluster is described by its n-dimensional mean vector. Whereas some modifications of the algorithm allow an adaptive fit of k to the data samples, the problem described above requires to set $k = 2$, as the task is to find two separate distributions of samples. Unfortunately, using exactly two clusters brings up a weakness of this method concerning outliers. In order to prevent identifying outliers as a cluster, it is recommended to remove outliers prior to the analysis, e.g. by using the 3σ rule, i.e. eliminating samples which deviate from the mean value by more than three times the standard deviation, compare (Wadsworth, 1997).

Having found two clusters, C_0 and C_1, the transformation routine $P_C : \mathbb{R} \to \{0,1\}$ can be defined as follows and the time series $x_i \in \mathbb{R}$ can be transformed to $y_i \in \{0,1\}$

$$y_i = \begin{cases} 0 & x \in C_0 \\ 1 & x \in C_1 \end{cases} \tag{11}$$

Just as the design of the transformation routine depends on the task considered, certain parameters have to be set depending on it. Since the task in this case is to detect a clustering of samples from different distributions, it is recommended to set the length of the analyzed subsequence τ in section 2.1 equals

$$\tau = min\{|C_0|, |C_1|\} \tag{12}$$

by default. In order to reduce the number of false alarms it is helpful to define an offset. This has the effect that a time series can only be reduced to a certain minimum number of samples. Another strategy to prevent false alarms is to demand a minimum size of each cluster. Both points are justified by the goal to analyze whether the distribution of samples has reliably changed and choice of settings should depend on risks associated with increasing either type I or type II error.

2.3 Dealing with Multiple Structural Breaks

In order to deal with multiple structural breaks, an iterative procedure of the algorithm presented within this paper is applied. Given a time series $x_1, \ldots, x_s$ and the algorithm detects a structural break at time index b, the algorithm is applied again on time series $x_b, \ldots, x_s$ until convergence, i.e. no further change-point is detected on the subsequence. If one is interested in identifying all change-points, the procedure can be applied to all remaining subsequence until convergence.

2.4 A Brief Note on Dealing with Nonadditive Changes

Nonadditive changes are defined in e.g. (Basseville and Nikiforov, 1993) as changes in variance, correlations, spectral characteristics, and dynamics of the signal or system. Hence, these types of changes are considered to be more complex to detect than additive changes, i.e. shifts in the mean value. Although additive changes play the central role in the following application on real data, the algorithm can easily be adapted to detect nonadditive changes. In order to demonstrate the flexibility of the novel approach, a rough recipe for this adaptation is provided.

The task of detecting either additive or nonadditive changes can be summarized as generating residuals of the original measurements that reflect the changes of interest, which are in this particular case of nonadditive nature. As stated above, instead of residuals the algorithm introduced within this paper demands a sequence of binary numbers, which is interpreted as the outcome of a binary stochastic process $\{Y_i\}_{i\in\mathbb{N}}$ with $\Omega := \{0,1\}$. Afterwards, the se-

quence can be analyzed using Chernoff's Bounding Method as described in section 2.1.

Alternatively to the transformation routine P_C, introduced in section 2.2, one can define new routines to face nonadditive changes. Specifically when analyzing changes in variance or higher statistical moments, one challenge is to avoid problems with shifting means in time series. Therefore, preprocessing in terms of e.g. high pass or wavelet filtering is recommendable, of which (Strang, 1989) provides a good survey on the latter. The outcome of the preprocessing shall be denoted as $x'_1,\ldots,x'_s \in \mathbb{R}$ and is assumed to be free of shifts in mean.

In a second step, the following transformation results in a reduction of the variance change detection problem to the additive change detection problem solved by the procedure defined in section 2.2.

$$\dot{x}_i = \begin{cases} \|x'_1\| & i = 1 \\ \left\|x'_i - x'_{i-1}\right\| & i \geq 2 \end{cases} \tag{13}$$

Assuming that elements of time series $x_1,\ldots,x_s$ are stochastically independent and that elements x_i and x_{i+1} follow the same distribution, it is known that the variance of distributions of derivatives of two i.i.d. variables summarizes to $2\cdot\sigma^2$, compare (Feller, 2009). However, this ensures that information on shifts in variance is not destroyed by the derivation in equation 13. Furthermore, using the absolute value in equation 13 and the symmetric character of the derivatives distribution reduces the problem to the additive change detection problem.

3 APPLICATION ON REAL DATA

This section provides a real-world application of the algorithm presented in this paper. The evaluation of the algorithm is based on 8002 real-world time series of a pharmaceutical wholesaler. These time series represent historical demands and, in their very nature, can imply seasonality, trends, slow or fast moving articles, or nonadditive changes as well as additive changes. The elements of each time series will be considered as independent and of unknown distribution, since no *a priori* information is available. Goal of this section is to show that the detection and removal of additive changes using the novel approach will reduce the forecast error significantly.

In section 3.1 forecast methods used for this evaluation are shortly introduced. In order to compare the novel approach to competitive strategies test scenarios are defined in section 3.2. Furthermore, the relative forecast error is defined as a measure to compare two given strategies. In section 3.3 results of evaluation are presented and discussed.

3.1 Forecast Methods

In order to estimate the value of preprocessing the following forecast methods have been implemented and applied on original and shortened time series.

- The arithmetic mean estimator is used as a representative of naive forecasting procedures. Additionally, this estimator should perform well on stationary time series.
- Single exponential smoothing is considered to be robust on seasonality, seasonal correlation, changing trends and suitable for forecasting in the presence of outliers as quoted in (Taylor, 2010) and (Gelper et al., 2010). Therefore, it should perform well even in the presence of structural changes. Considered for original work are Brown and Holt in the 1950s, compare e.g. (Holt, 1957) and (Brown, 1959), and a review on exponential smoothing in general is provided in (Gardner Jr, 1985).
- Linear regression analysis is recommendable for predictions on basis of time series containing trends. State of the art applications are provided in (Ng et al., 2008), (Xia and Zhao, 2009), and (Pinson et al., 2008) to name but a few.

In combination, these algorithms address important issues of time series prediction. The selection procedure to decide which forecast method should be used for a particular time series can be described as best historical performance principle. This principle pretends that historical performance is an indicator for future performance. Formally speaking, the goal is to determine a method to predict $\hat{x}_{s+1}$. Each forecast method available can now be used to forecast w samples $\hat{x}_{s-w+1},\ldots,\hat{x}_s$ of time series $x_1,\ldots,x_s$. The best method is determined e.g. with respect to the average mean squared error

$$AMSE = \frac{1}{w}\sum_{t=s-w+1}^{s}(\hat{x}_t - x_t)^2 \tag{14}$$

and used to estimate $\hat{x}_{s+1}$.

3.2 Design of Test Scenarios and Relative Error Estimates

The goal is to analyze whether preprocessing in terms of structural break detection is an improvement to forecasting or not. Hence, test scenarios will be defined which are composed of two preprocessing

Table 1: Overview on scenarios. Each scenario contains a reviewed preprocessing strategy, a reference preprocessing strategy and a set of forecast functions applied on either preprocessed time series.

Scenario ID	Reviewed Strategy	Reference Strategy	Forecast Functions
1	CB	None	AME
2	CB	NA	AME
3	CB	BinDist	AME
4	CB	None	Combo
5	CB	NA	Combo
6	CB	BinDist	Combo
7	BinDist	None	AME
8	BinDist	NA	AME
9	BinDist	CB	AME
10	BinDist	None	Combo
11	BinDist	NA	Combo
12	BinDist	CB	Combo

modes and a set of forecast functions. The preprocessing mode can be any one of the following:

- None (None). No preprocessing in the sense of structural break detection is applied at all. This mode will be used to illustrate the value of structural break detection.
- Chernoff Bounds (CB). The algorithm presented in this paper is applied for structural break detection. If a break is detected, the time series is abridged accordingly.
- Binomial distribution (BinDist). The algorithm presented in this paper is applied for structural break detection but instead of Chernoff's approximations the exact bounds of the Binomial distribution are used. This is done for comparison of both thresholds. Previous work (Pauli et al., 2011) has shown that for short time series containing no more than 150 samples, the run time of the algorithm using exact thresholds rather than Chernoff Bounds can be approximated by a factor of three.
- Naive approach (NA). In order to compare accurate detection methods to a naive approach, one strategy will be to cut off the time series at point $b = \lceil s/2 \rceil$.

Furthermore, two sets of forecasting functions are defined:

- The first set (AME) only contains the arithmetic mean estimator. This is reasonable since it represents naive forecasting methods and should perform well especially on stationary time series.
- The second set (Combo) contains the arithmetic mean estimator, single exponential smoothing and linear regression for reasons given in section 3.1.

Scenarios are composed of preprocessing strategies and a set of forecast functions. Table 1 provides a list of all scenarios to be evaluated in section 3.3. The relative forecast error is estimated for each time series separately in the following way. In order to reduce type II errors or false alarms, the best historical performance principle, as introduced in section 3.1 for forecasting, is applied for the selection of the preprocessing strategy as well. If the reviewed strategy performed better in the past on $x_1, \ldots, x_{s-1}$ than the reference strategy in terms of *AMSE*, then the reviewed strategy is used for the actual forecast of sample x_s as well. If the reviewed strategy performed better in the past, then the relative error is measured. The *AMSE* received by the reference strategy is denoted as $AMSE^{Ref}$ and the *AMSE* received by the reviewed strategy as $AMSE^{Rev}$ and the estimates $\hat{x}_s^{Ref}$ and $\hat{x}_s^{Rev}$ at time index s, respectively. The residua are denoted as δ_s^{Ref} and δ_s^{Rev}, respectively.

Then the relative error R_η of time series η is defined as

$$R_\eta := \begin{cases} \frac{\left|\delta_s^{Ref}\right| - \left|\delta_s^{Rev}\right|}{\left|\delta_s^{Ref}\right|} & \left|\delta_s^{Rev}\right| < \left|\delta_s^{Ref}\right| \\ \frac{\left|\delta_s^{Ref}\right| - \left|\delta_s^{Rev}\right|}{\left|\delta_s^{Rev}\right|} & \left|\delta_s^{Rev}\right| > \left|\delta_s^{Ref}\right| \\ 0 & \text{else} \end{cases} \quad (15)$$

Consider that the *AMSE* is estimated on $x_1, \ldots, x_{s-1}$ and the improvement might be negative, if $\hat{x}_s^{Rev}$ proves to be a worse estimator than $\hat{x}_s^{Ref}$, even if $AMSE^{Rev} < AMSE^{Ref}$. Hence, missed and false alarms will be measured as described in table 2. Whereas missed structural breaks fail to reduce forecast errors, false structural breaks increase forecast errors. Obviously it is worthwhile avoiding both of them. Formula 15 returns the percentage error decrease in case $\hat{x}_s^{Rev}$ is a better estimator than $\hat{x}_s^{Ref}$ and the percentage error increases in case of false alarms.

Finally, the relative error improvement E_η of time series η is defined as

$$E_\eta := \begin{cases} R_\eta & AMSE^{Rev} < AMSE^{Ref} \\ 0 & \text{else} \end{cases} \quad (16)$$

3.3 Evaluation

The evaluation of the algorithm is based on 8002 real-world time series of a pharmaceutical wholesaler. The elements of each time series have been considered to be independent and of unknown distribution. In this section, results of test scenarios as defined in section

Table 2: Classification of historical and present strategies. In order to reduce false alarms the best historical performance principle is applied, but on account of missed alarms. Measuring false and missed alarms indicates success of the method.

Classification	Historical Performance	Present Performance
Sensitivity	$AMSE^{Rev} > AMSE^{Ref}$	$\delta_s^{Rev} > \delta_s^{Ref}$
Specificity	$AMSE^{Rev} < AMSE^{Ref}$	$\delta_s^{Rev} < \delta_s^{Ref}$
False alarm	$AMSE^{Rev} < AMSE^{Ref}$	$\delta_s^{Rev} > \delta_s^{Ref}$
Missed alarm	$AMSE^{Rev} > AMSE^{Ref}$	$\delta_s^{Rev} < \delta_s^{Ref}$

3.2 are discussed. In order to increase the clarity of graphical presentation, the scenarios have been subdivided into four groups of three each. The performance measures in terms of significance, power, forecast error improvement and runtime are summarized in table 3 for all scenarios.

When performing the tests, it became obvious that results have been volatile to a certain extend for the following reason. Assume $AMSE^{Rev} < AMSE^{Ref}$ and $\hat{x}^{Rev}$ is taken as the next forecast, ϕ is the true distribution of $x_s \in X$ and

$$\left|E[X] - \hat{x}_s^{Rev}\right| < \left|E[X] - \hat{x}_s^{Ref}\right| \tag{17}$$

then the probability P that $\hat{x}_s^{Ref}$ is a better estimator for x_s than $\hat{x}_s^{Rev}$ is given by

$$P := \begin{cases} \int_a^{\infty} \phi(x)\,dx & \hat{x}_s^{Rev} < \hat{x}_s^{Ref} \\ \int_0^{a} \phi(x)\,dx & \hat{x}_s^{Rev} > \hat{x}_s^{Ref} \end{cases} \tag{18}$$

where $a = \frac{1}{2}\left(\hat{x}_s^{Rev} + \hat{x}_s^{Ref}\right)$.

In order to reduce the volatility of the results, the test sequence to determine the performance indicators has been increased from one to ten, or in other words, instead of estimating $\hat{x}_s^{Rev}$ and $\hat{x}_s^{Ref}$, the sequences $\hat{x}_{s-9}^{Rev}, \ldots, \hat{x}_s^{Rev}$ and $\hat{x}_{s-9}^{Ref}, \ldots, \hat{x}_s^{Ref}$ have been estimated. Figure 2 shows the cdf for the first three scenarios. In each of the three scenarios, structural break detection using Chernoff Bounds was the reviewed method and the arithmetic mean estimator was the only forecast method used. As can be seen in the figure, the Chernoff Bounds competed well against all three competitors. The cdf takes a forecast error into account if an additive change has been detected and therefore $AMSE^{Ref} \neq AMSE^{Ref}$. The relative error has been measured as depicted in equations 16 and 15. The curve shows that for scenario one about 40% of the forecasts could be improved if a structural break had been detected. In 18% of the forecasts, the error could be reduced by more than 50%. The forecast error was increased by 50% in less than 2% of the forecasts, due to false alarms. The first scenarios show that especially when dealing with naive forecast methods, structural break detection results in great improvements in terms of relative forecast errors.

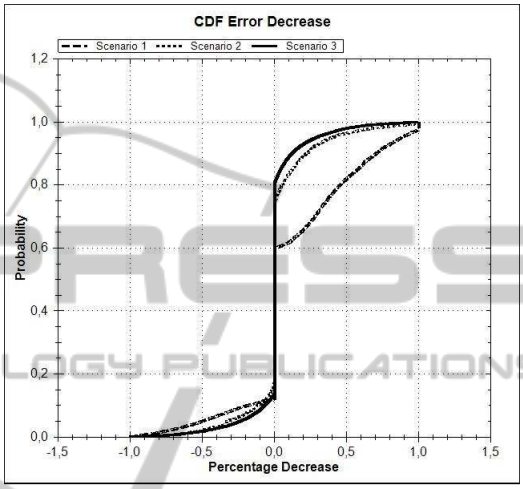

Figure 2: CDF's of the relative forecast error reduction of scenario one, two and three.

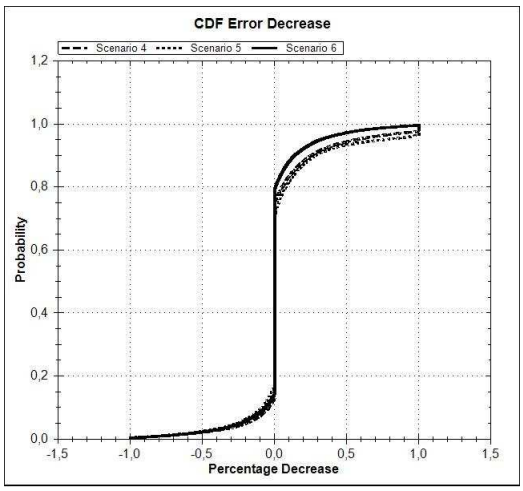

Figure 3: CDF's of the relative forecast error reduction of scenario four, five and six.

Figure 3 displays similar scenarios to those shown in figure 2, with more sophisticated forecast methods having been used in the former. Comparing scenario one and four, the effect of improving forecast methods can be seen if no preprocessing has been applied

Table 3: Sensitivity, missed alarm, specificity, and false alarm classify success in terms of forecast error reduction only if a structural break has been detected. Ratios of improved or worsened forecasts reflect success in terms of forecast error reduction proportional to the overall number of forecasts done. The runtime is standardized by the fastest scenario, which took approximately four seconds.

Scenario ID	Sensitivity	Missed Alarm	Specificity	False Alarm	Ratio of Improved FCs	Ratio of Worsen FCs	Relative Runtime
1	0.67	0.33	0.81	0.19	0.26	0.06	1
2	0.71	0.29	0.62	0.38	0.28	0.17	1
3	0.78	0.22	0.71	0.29	0.21	0.09	3
4	0.71	0.29	0.73	0.27	0.20	0.07	190
5	0.62	0.38	0.62	0.38	0.29	0.18	96
6	0.77	0.23	0.72	0.28	0.25	0.10	103
7	0.61	0.39	0.79	0.21	0.33	0.09	2
8	0.67	0.33	0.65	0.35	0.30	0.16	2
9	0.71	0.29	0.78	0.22	0.26	0.07	3
10	0.64	0.36	0.71	0.29	0.24	0.10	127
11	0.60	0.40	0.64	0.36	0.28	0.16	82
12	0.72	0.28	0.77	0.23	0.24	0.07	103

before. Scenarios one to six have been repeated, using the exact bounds of the binomial distribution instead of Chernoff's approximations. The results of scenario seven to twelve, compare figure 4, are similar to those of scenario one to six, but using the exact bounds increases the relative forecast error reduction as expected.

Table 3 extends the results given in figures 2 to 4. As defined in table 2, sensitivity, specificity, false alarms, and missed alarms have been measured. This classification can only take into account forecasts of time series, for which structural breaks have been detected. Since the time series represent real-life data instead of artificial ones, break dates are unknown. Therefore relative error reduction is used as performance measure. The ratio of improved or worsened forecasts takes all forecasts into account, i.e. it is the absolute number of specificities or false alarms divided by the total number of forecasts done, respectively. For example in scenario one, 26% of all forecasts have been positively influenced by using structural break detection.

The results in table 3 show the positive effect of using sophisticated preprocessing methods and diverse forecasting methods. Results of scenarios in which naive preprocessing was involved appear to be arbitrary, indicated by high ratios of both improved and worsened forecasts.

The runtime of scenarios has been standardized. Scenario one took about four seconds to complete. Previous studies in (Pauli et al., 2011) show that the runtime of exact bounds differs by a factor of three in comparison to Chernoff's Inequalities for short time series and increases exponentially for longer time series. The increase of runtime when using a combination of forecasting methods is considerable. Taking into account runtime and error reduction ratios, applying preprocessing methods appears to be very worthwhile.

4 CONCLUSIONS AND FUTURE PROSPECTS

In section 3.3 the evaluation of the novel approach to structural break detection and its impact on forecasting have been performed and discussed. Results are striking in terms of forecast error reduction and runtime as can be seen in table 3.

The scenarios contained both naive and sophisticated forecast and structural break detection methods. Table 3 shows that using sophisticated forecast methods raises the runtime enormously in relation to its error improvement, compare scenario one and four. A possible explanation for the success of preprocessing in terms of change-point detection might be that it is more widely applicable than additional forecast algorithms. New forecasting algorithms are often designed to deal with special characteristics on certain time series, whereas prepocessing will affect forecasting performance in a wider range of time series.

The algorithm used in this paper is designed to deal with additive changes, i.e. shifts in the mean value. Nonadditive changes which occur in variance, correlations, spectral characteristics, and dynamics of the signal or system, compare (Basseville and Nikiforov, 1993), are the topic of future work. It will be

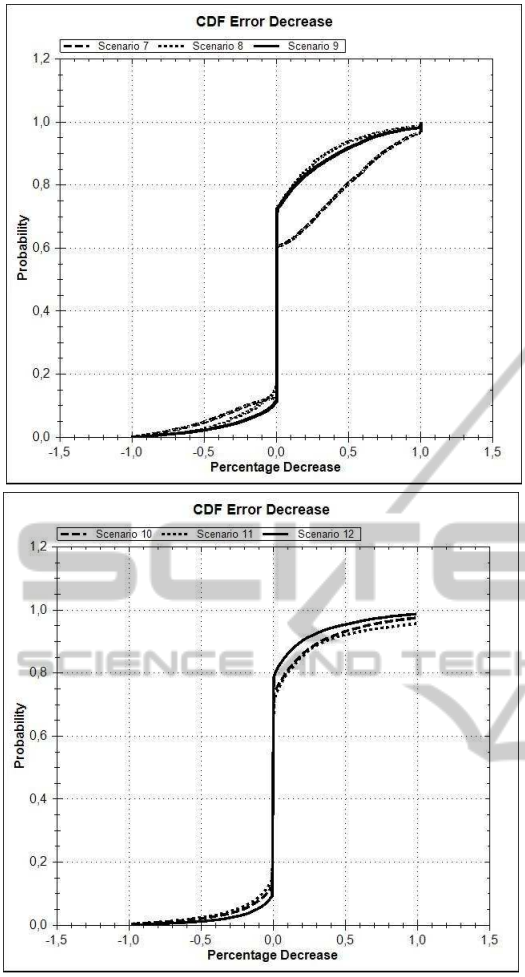

Figure 4: CDF's of the relative forecast error reduction of scenario seven to twelve.

shown that this detection problem can be solved by adequate transformation routines that reflect changes of interest. The special aim will be to design those transformation routines with respect to efficiency and robustness. Section 2.4 provided a brief note to this topic in order to show the flexibility of the novel approach.

From the theoretical point of view the current application is an offline detection problem. In future work this algorithm will be applied to online detection problems, which demands new performance indicators such as mean time between false alarms or mean delay for detections.

The goal of this paper is to demonstrate the applicability of the algorithm to a real-world problem and facing real-world data. Future prospects will be to analyze more general performance indicators as proposed for example in (Basseville and Nikiforov, 1993) such as mean time between false alarms, probability of false detections, mean delay for detection, probability of nondetection, statistical power, and required effect size to name but a few. The goal will be to answer these questions analytically and by simulation.

REFERENCES

Basseville, M. and Nikiforov, I. (1993). *Detection of Abrupt Changes: Theory and Application*. Prentice-Hall,Inc.

Brown, R. (1959). *Statistical forecasting for inventory control*. McGraw-Hill New York.

Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, 41(3):1–58.

Chernoff, H. (1952). A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the Sum of Observations. *The Annals of Mathematical Statistics*, 23:493–507.

Feller, S., Chevalier, R., and Morsili, S. (2010). Parameter Disaggregation for High Dimensional Time Series Data on the Example of a Gas Turbine. In *Proceedings of the 38th ESReDA Seminar, Pcs, H*, pages 13–26.

Feller, W. (2009). *An introduction to probability theory and its applications*. Wiley-India.

Fujimaki, R., Yairi, T., and Machida, K. (2005). An Approach to Spacecraft Anomaly Detection Problem Using Kernel Feature Space. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pages 401–410. ACM.

Gardner Jr, E. (1985). Exponential smoothing: The state of the art. *Journal of Forecasting*, 4(1):1–28.

Gelper, S., Fried, R., and Croux, C. (2010). Robust forecasting with exponential and Holt-Winters smoothing. *Journal of Forecasting*, 29(3):285–300.

Guralnik, V. and Srivastava, J. (1999). Event Detection from Time Series Data. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 33–42. ACM.

Gustafsson, F. (1998). Estimation and Change Detection of Tire-Road Friction Using the Wheel Slip. *IEEE Control System Magazine*, 18(4):42–49.

Hartigan, J. and Wong, M. (1979). Algorithm AS 136: A k-means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108.

Holt, C. (1957). Forecasting trends and seasonals by exponentially weighted moving averages. *ONR Memorandum*, 52:1957.

Ibaida, A., Khalil, I., and Sufi, F. (2010). Cardiac abnormalities detection from compressed ECG in wireless telemonitoring using principal components analysis (PCA). In *Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), 2009 5th International Conference on*, pages 207–212. IEEE.

Ide, T. and Kashima, H. (2004). Eigenspace-based Anomaly Detection in Computer Systems. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 440–449. ACM.

Kawahara, Y. and Sugiyama, M. (2009). Change-point Detection in Time Series Data by Direct Density-Ratio Estimation. In *Proceedings of 2009 SIAM International Conference on Data Mining (SDM2009)*, pages 389–400.

Ma, J. and Perkins, S. (2003). Time Series Novelty Detection Using One-class Support Vector Machines. In *Proceedings of the International Joint Conference on Neural Networks*, volume 3, pages 1741–1745.

Markou, M. and Singh, S. (2003). Novelty Detection: a Review–Part 1: Statistical Approaches. *Signal Processing*, 83(12):2481–2497.

Murad, U. and Pinkas, G. (1999). Unsupervised Profiling for Identifying Superimposed Fraud. *Principles of Data Mining and Knowledge Discovery*, 1704:251–261.

Ng, T., Skitmore, M., and Wong, K. (2008). Using genetic algorithms and linear regression analysis for private housing demand forecast. *Building and Environment*, 43(6):1171–1184.

Pauli, D., Timm, I., Lorion, Y., and Feller, S. (2011). Using Chernoff's Bounding Method for High-Performance Structural Break Detection. Submitted for publication.

Perron, P. (2006). Dealing with Structural Breaks. *Palgrave handbook of econometrics*, 1:278–352.

Pinson, P., Nielsen, H., Madsen, H., and Nielsen, T. (2008). Local linear regression with adaptive orthogonal fitting for the wind power application. *Statistics and Computing*, 18(1):59–71.

Press, W., Teukolsky, S., Vetterling, W., and Flannery, B. (2007). *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press.

Schwabacher, M., Oza, N., and Matthews, B. (2007). Unsupervised Anomaly Detection for Liquid-Fueled Rocket Propulsion Health Monitoring. In *Proceedings of the AIAA Infotech@ Aerospace Conference, Reston, VA: American Institute for Aeronautics and Astronautics, Inc*.

Strang, G. (1989). Wavelets and dilation equations: A brief introduction. *Siam Review*, 31(4):614–627.

Taylor, J. (2010). Multi-item sales forecasting with total and split exponential smoothing. *Journal of the Operational Research Society*.

Wadsworth, H. (1997). *Handbook of statistical methods for engineers and scientists*. McGraw-Hill Professional.

Xia, B. and Zhao, C. (2009). The Application of Multiple Regression Analysis Forecast in Economical Forecast: The Demand Forecast of Our Country Industry Lavation Machinery in the Year of 2008 and 2009. In *Second International Workshop on Knowledge Discovery and Data Mining, 2009. WKDD 2009*, pages 405–408.